Embedding Model Choice in RAG Pipelines: A Multi-Metric Comparison of all-mpnet-base-v2 and all-MiniLM-L6-v2

Ricci Doddz Juaman

Abstract

Retrieval-Augmented Generation (RAG) systems rely critically on the quality of their embedding models to retrieve relevant context for downstream language models. While high-capacity embedding models such as all-mpnet-base-v2 are often assumed to deliver superior performance, smaller and more efficient models such as all-MiniLM-L6-v2 remain attractive for production use due to latency and cost considerations. This paper presents a controlled, multi-metric evaluation of these two sentence embedding models under both semantic-only and hybrid semantic-lexical retrieval configurations. Using a high-relevance-density Gold Set and evaluating Recall@k, Precision@k, Mean Reciprocal Rank (MRR), and nDCG at cutoffs of 5, 10, and 15, we analyze not only retrieval completeness but also ranking efficiency and context purity. Results show that while all-mpnet-base-v2 benefits substantially from hybrid lexical augmentation, all-MiniLM-L6-v2 exhibits strong ranking stability in semantic-only settings, with hybrid retrieval introducing mixed tradeoffs. These findings highlight that embedding model choice in RAG pipelines should be guided by ranking behavior and system constraints rather than embedding capacity alone.

---

1. Introduction

The rapid adoption of Retrieval-Augmented Generation systems has shifted the optimization problem from prompt engineering to retrieval quality. In modern RAG pipelines, the embedding model serves as the primary bottleneck between user intent and the language model's ability to generate accurate responses. Despite this, embedding model selection is often treated as a secondary concern, with limited empirical evaluation beyond recall-oriented benchmarks.

This work investigates how embedding model choice impacts retrieval behavior in a RAG context. Specifically, we compare all-mpnet-base-v2, a high-capacity sentence transformer commonly used as a benchmarking standard, with all-MiniLM-L6-v2, a lighter-weight model optimized for efficiency. Rather than evaluating these models in isolation, we examine their

interaction with a hybrid retrieval strategy that combines dense semantic similarity with BM25 lexical scoring.

Our goal is not to identify a universally “best” embedding model, but to characterize tradeoffs that emerge across multiple evaluation dimensions, particularly those most relevant to RAG systems: ranking efficiency, precision under constrained context windows, and robustness to noisy candidate pools.

---

2. System Architecture

2.1 Embedding Models

all-mpnet-base-v2
all-mpnet-base-v2 is a transformer-based sentence embedding model widely regarded as a high-performance baseline in semantic retrieval tasks. With its larger parameter count and richer representational capacity, it excels at capturing nuanced semantic relationships between queries and documents.

all-MiniLM-L6-v2
all-MiniLM-L6-v2 is a compact embedding model designed for efficiency and speed. Despite its reduced size, it has demonstrated competitive performance in many retrieval benchmarks, making it a strong candidate for production RAG systems where latency and cost constraints are significant.

2.2 Lexical Scoring and Hybrid Retrieval

To complement semantic embeddings, BM25 lexical scoring was incorporated as a weighted component in a hybrid retrieval configuration. The hybrid score is defined as a linear combination of normalized semantic similarity and BM25 relevance, with a fixed lexical weight of 0.3.

The motivation for hybrid retrieval lies in addressing semantic drift, where dense embeddings retrieve topically related but lexically mismatched documents. Lexical scoring provides a corrective signal that emphasizes exact term matches, particularly for queries involving explicit keywords such as “refund” or “delivery issue.”

---

3. Experimental Setup

3.1 Gold Set Construction

To enable controlled and repeatable evaluation, a high-relevance-density Gold Set was constructed. Seven distinct queries were selected, spanning product-related, experiential, and service-oriented intents. For each query, a candidate pool of 100 documents was assembled, consisting of 30 relevant and 70 irrelevant documents.

This 30% relevance density intentionally emphasizes ranking discrimination over coarse recall. In such settings, small ranking errors are amplified, making metrics such as Precision@k, MRR, and nDCG particularly informative.

3.2 Evaluation Metrics

Retrieval performance was evaluated using the following metrics at cutoffs of 5, 10, and 15:

- **Recall@k**: Measures the proportion of relevant documents retrieved.
- **Precision@k**: Measures the purity of the retrieved set.
- **MRR@k**: Captures how early the first relevant document appears.
- **nDCG@k**: Evaluates ranking quality while accounting for graded relevance and position bias.

Mean values across all queries were used to assess overall system behavior.

---

4. Results and Analysis

4.1 all-mpnet-base-v2: Semantic vs Hybrid Retrieval

For all-mpnet-base-v2, hybrid retrieval consistently improved performance across nearly all metrics.

- **Recall@15** increased from 0.208 to 0.222.
- **Precision@15** improved from 0.467 to 0.495.
- **MRR@15** showed a substantial gain from 0.833 to 0.929.

- **nDCG@15** increased from 0.517 to 0.561.

These improvements indicate that lexical augmentation not only increases retrieval completeness but significantly enhances ranking efficiency. The most relevant documents are pushed closer to the top of the result set, resulting in a cleaner and more informative context window for downstream language models.

4.2 all-MiniLM-L6-v2: Semantic vs Hybrid Retrieval

In contrast, all-MiniLM-L6-v2 exhibited a different performance profile.

- **Recall@15** showed a marginal increase from 0.219 to 0.224 under hybrid retrieval.
- **Precision@15** improved only slightly, from 0.438 to 0.448.
- **MRR@15**, however, decreased from 1.000 to 0.905.
- **nDCG@15** increased from 0.783 to 0.834.

These results suggest that while lexical scoring improves overall ranking distribution (as reflected by nDCG), it can negatively impact early-rank stability for MiniLM. In several cases, hybrid retrieval displaced highly ranked semantic matches, resulting in lower MRR despite modest gains in recall and precision.

4.3 Cross-Model Comparison

When comparing the two models directly, several patterns emerge:

- all-mpnet-base-v2 benefits strongly from hybrid retrieval, particularly in MRR and nDCG.
- all-MiniLM-L6-v2 demonstrates strong semantic-only ranking stability, often achieving perfect MRR in baseline configurations.
- Hybrid retrieval introduces greater variability for MiniLM, suggesting sensitivity to lexical weighting.

These findings indicate that larger embedding models may rely more heavily on lexical correction to refine rankings, whereas smaller models may already encode stronger lexical-semantic alignment in their dense representations.

---

5. Discussion

The results challenge the assumption that larger embedding models are universally superior in RAG pipelines. While all-mpnet-base-v2 achieves higher peak performance when combined with lexical scoring, all-MiniLM-L6-v2 delivers competitive and sometimes more stable results in semantic-only configurations.

From a systems perspective, this has practical implications:

- For latency-sensitive pipelines, MiniLM may offer sufficient retrieval quality without hybrid complexity.
- For precision-critical applications, mpnet with hybrid retrieval provides more robust ranking guarantees.
- Embedding model choice should be evaluated jointly with retrieval strategy, not in isolation.

---

6. Conclusion

This study demonstrates that embedding model selection in RAG pipelines is a multi-dimensional decision involving tradeoffs between recall, precision, ranking efficiency, and system constraints. Through a controlled, high-density evaluation, we show that all-mpnet-base-v2 and all-MiniLM-L6-v2 exhibit distinct retrieval behaviors that are further shaped by the inclusion of lexical scoring.

Rather than converging on a single “best” embedding model, effective RAG system design requires aligning model capacity, retrieval strategy, and evaluation metrics with the intended application. Multi-metric benchmarking, particularly emphasizing MRR and nDCG, is essential for understanding these tradeoffs and building reliable retrieval systems.